Thoughtful Precision in Mini-Apps

[1,2]Siddhartha Bishnu and [1,3]Shane Fogerty

[1]*Los Alamos National Laboratory,* [2]*Florida State University,* [3]*University of Rochester*

LA-UR-17-26611

Abstract

- Approximate computing can address a lot of challenges in exascale computing.
- We studied approximate approaches to solving a range of Department of Energy (DOE) relevant computational problems on a variety of architectures.
- Anticipated **improvements** are observed in **computational** and **memory performance** as well as in **power savings**.
- **Application correctness** is determined to be **within acceptable bounds** while operating under the conditions of reduced precision.

Background

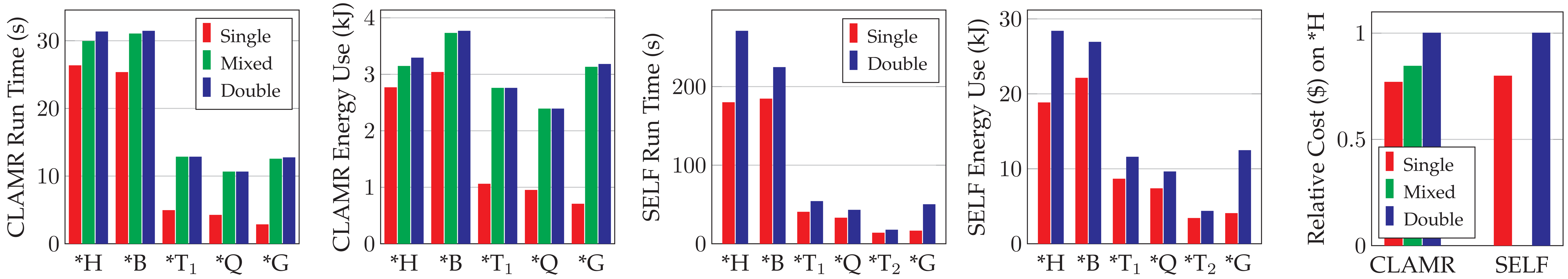
- Floating-point numbers can consist of 16-bits (half precision), 32-bits (single precision), 64-bits (double precision) etc.
- Mixed precision sets the large physical state arrays to single precision, but promotes all local calculations to double precision.
- Mixed-precision code can sometimes achieve **similar accuracy** to its double-precision counterpart while being **significantly faster and reducing memory pressure**.
- Instead of reducing precision everywhere, it is advisable to focus on **choosing the level of precision according to the needs of the calculation**, to the extent of increasing precision in well-chosen sub-calculations e.g. **global sums** and lowering it elsewhere.

Methodology

We investigated the impacts of varying precision on **CLAMR**, a hydrodynamic cell-based adaptive mesh refinement DOE mini-app and another mini-app **SELF** (Spectral Element Libraries in Fortran) on different architectures viz.

- Intel Xeon E5-2660_v3 **Haswell (*H)** CPU
- Intel Xeon E5-2695_v4 **Broadwell (*B)** CPU
- Nvidia **Tesla K40m (*T_1)** GPU
- Nvidia **Quadro K6000 (*Q)** GPU
- Nvidia **Tesla P100 (*T_1)** GPU
- Nvidia **GeForce GTX TITANX (*G)** GPU

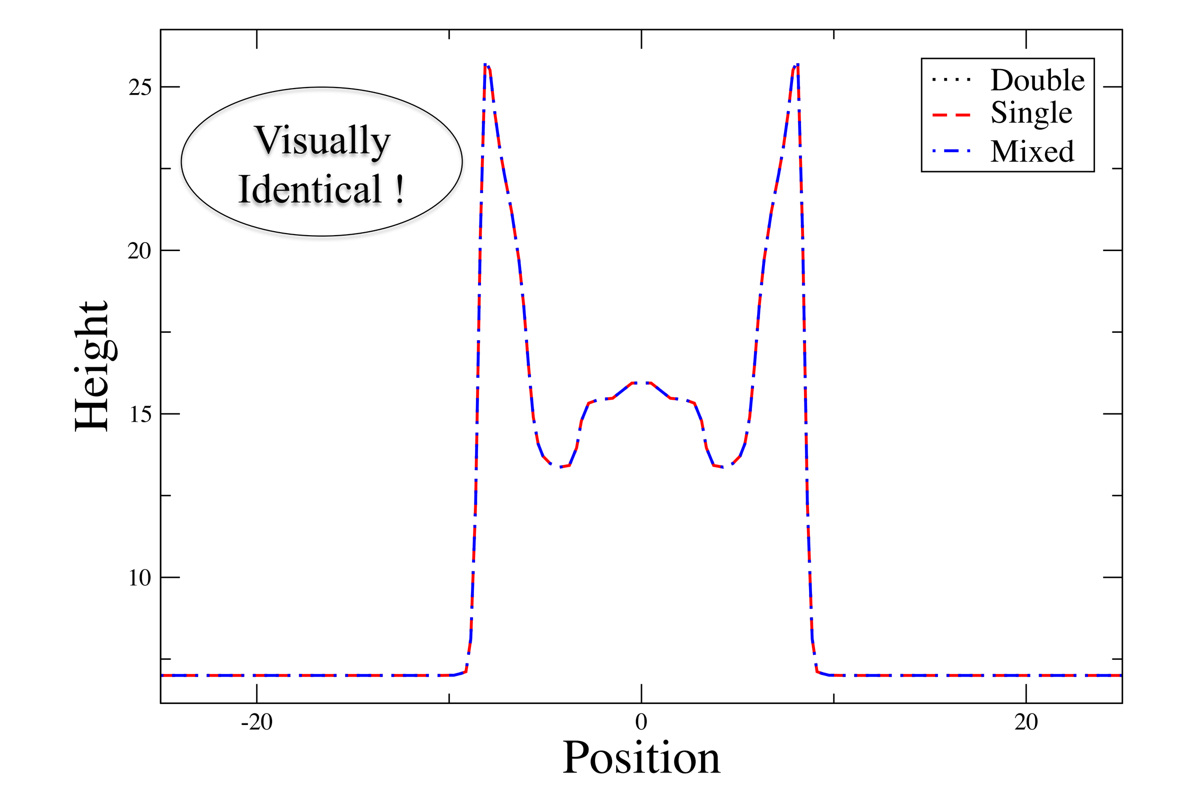
Precision Cost Analysis Results



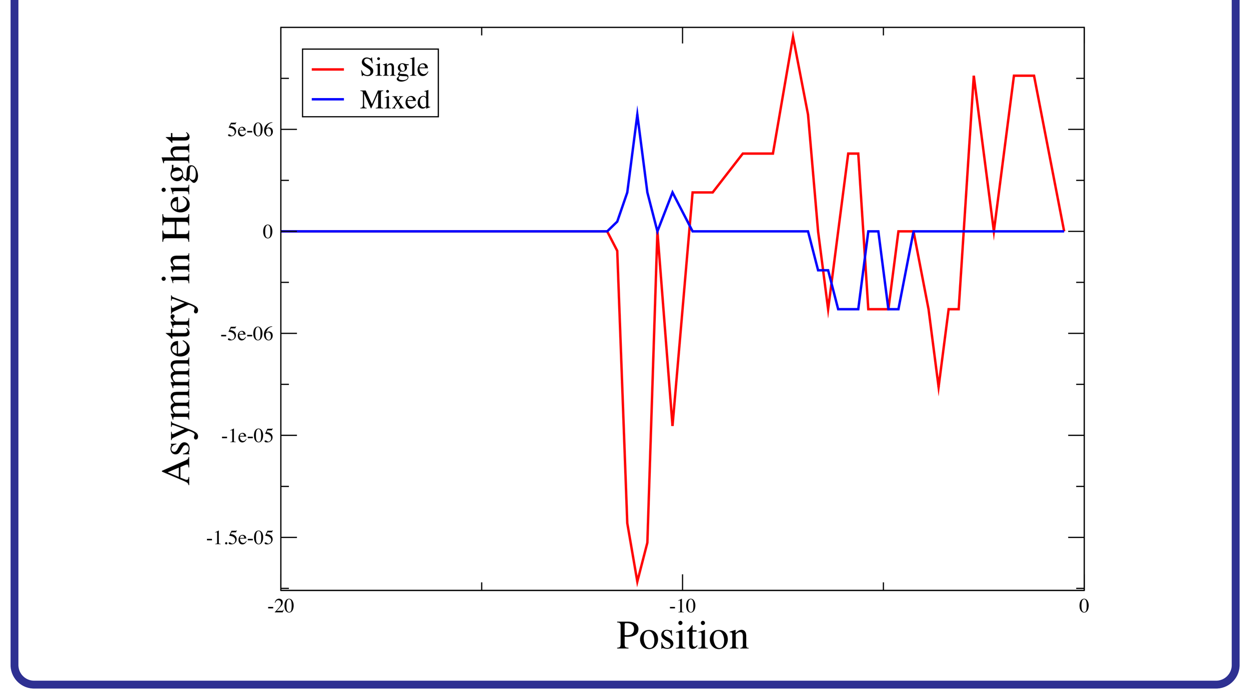
*H ≡ Haswell, *B ≡ Broadwell, *T_1 ≡ Tesla K40m, *Q ≡ Quadro K6000, *T_2 ≡ Tesla P100 and *G ≡ GeForce GTX TITANX

Accuracy Analysis

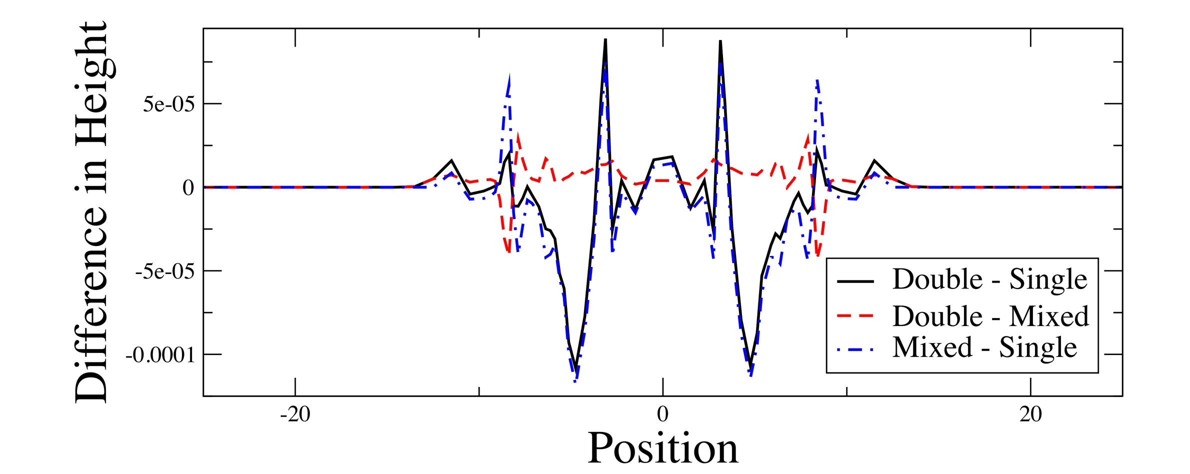
Slices of CLAMR simulation results with 64 × 64 grid points, 2 AMR levels and varying precision



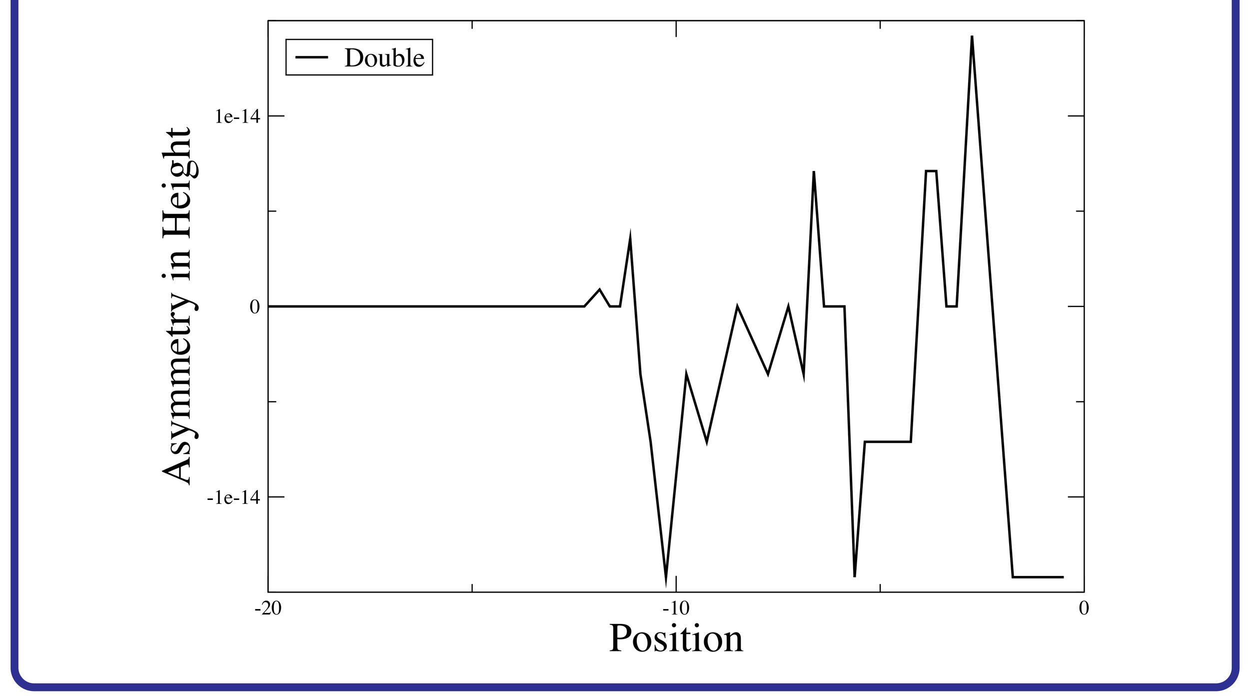
Height asymmetry for single & mixed precisions



Differences among the CLAMR simulations



Height asymmetry for double precision



Conclusion

- We have demonstrated that in two different DOE-relevant mini-applications, **reduced precision can save**
  - **computational time** and **cost**
  - **storage cost**
  - **memory use**
  - **power consumption**

  and **increase performance** significantly with **modest changes** to the application code base.
- **Careful implementation of precision** in well-chosen parts of the code can **preserve application correctness** to an appreciable degree.
- We can **complement lower precision** by **increasing** the **degrees of freedom**.
- **Hardware choice is important** as reduced precision greatly improves the performance of **CLAMR** on the **GPUs** and **SELF** on the **Haswell CPU** and the **GTX TITANX GPU**.
- This provides us with a great opportunity for hardware-software **codesign**.

It is time for application developers to jump on this disruptive trend in computing capabilities.

Acknowledgements

We thank the **Information Science and Technology Institute (ISTI)**, the **Parallel Computing Summer Research Internship (PCSRI)** program and the **Beyond Moore's Law project** at the **Los Alamos National Laboratory** for providing us with resources and support. This work was supported by the **Los Alamos National Laboratory** under contract **DE-AC52-06NA25396**.

References

[1] Shane Fogerty, Siddhartha Bishnu, Yuliana Zamora, Laura Monroe, Steve Poole, Michael Lam, Joe Schoonover, and Robert Robey. Thoughtful Precision in Mini-apps. Technical report, Los Alamos National Laboratory, 2017. LA-UR-17-25426.